BAYESIAN NETWORKS AND THE SEARCH FOR CAUSALITY

Ricardo Silva
ricardo@stats.ucl.ac.uk

Department of Statistical Science
Centre for Computational Statistics and Machine Learning

University College London
“I’d rather discover a single causal law than become the king of Persia.”
We Will

• Start with the very basics of causal inference

• Provide some basic background in Bayesian networks/graphical models

• Show how graphical models can be used in causal inference

• Describe application scenarios and the practical difficulties
What is a Causal Inference Problem?

Let me give you two problems.
Problem 1

You are in charge of setting the price of life insurance for a person you know is a smoker, among other things. What is your approach and what do you need to know?
Problem 2

You are in charge of public policy on smoking incentives. You want to minimise health costs that may be due to smoking. What is your approach and what do you need to know?
On Causation, Prediction and Explanation

- There are tasks of prediction, control and explanation.

- Prediction is bog-standard in machine learning, statistics, predictive analytics etc.

- Control is about taking actions to achieve a particular outcome.

- Explanation concerns what the outcome would be if you had seen different data. It involves actions that have not taken place.
Causal Inference

• Causal inference is essentially about control and explanation.

• Good control should require good predictive models anyway.

• Explanation is not about the future, but counterfactual events in the past.

• How to solve these problems?
Learning from Actions
Experimental Design

IV

AN AGRICULTURAL EXPERIMENT IN RANDOMISED BLOCKS

22. Description of the Experiment

In pursuance of the principles indicated in the discussions in the previous chapters we may now give an example from agricultural experimentation, a branch of the subject in which these principles are so far been most explicitly developed, and in which advantages and disadvantages of the different methods open to the experimenter may be most clearly displayed.

We will suppose that our experiment is designed to test the relative productivity, or yield, of five different varieties of a farm crop; and that a decision has been arrived at as to what produce shall be regarded as yield. In the case of cereal crops, for example, the yield may be measured as total grain, or as the grain sufficiently large to pass a specified sieve, or as grain and straw valued together at pre-determined prices, or in whatever manner as may be deemed appropriate for the purposes of the experiment. Our purpose is to determine whether, on the soil or in the climate experienced by the test, any of the various varieties have been better or worse than others, and, if so, to evaluate the degree of precision.
Experimental Design

• Say you have a choice of treatments, in order to understand a particular outcome.

• Along the line of Fisher’s examples, you could define as your outcome the productivity of a particular plantation field.

• As treatments, different combinations of fertilizers at different dosages.

• In the data, the choice of treatment is set by design, so we know how it was generated.
Exploitation of Findings

- Once we learn the relationship between treatment and outcome, we can use this information to come up with an optimal policy.
  - For instance, pick combination of fertilizers/dosage that maximises expected crop productivity.
  - This is essentially the application of decision theory.
Exploitation of Findings

• An alternative use is to understand what would have happened to those outcomes had treatment been different.
  • For instance, a marketing campaign was followed by major losses. How can we assign blame or responsibility for these outcomes?
  • This is an in-sample, NOT an out-of-sample estimand.

• This is essentially the application of counterfactual modelling.

• Notice: counterfactual analysis is NOT about prediction and control, which is my focus. For the rest of this talk, I’ll have little to say about counterfactual learning.
Interplay with Modelling

- The number of possible experimental conditions may explode, and treatment (action) levels can be continuous.

- All sorts of models (logistic regression, Gaussian processes etc.) can be used to map treatment to outcome.

- In particular, analysis of variance (ANOVA) via **Latin squares** is one of the most classical and practically used methods in some industries.
Interplay with Inference

• Traditional statistics techniques (power analysis, hypothesis testing, confidence intervals) are also used in experimental design.

• Fisher’s “The Design of Experiments” was one of the sources responsible (to blame?) for the popularity of hypothesis testing.

“0.05”
(Not really. Fisher knew better than that.)
Sidenote: A/B Testing and Bandits

- A/B testing is the baby sibling of experimental design.

- Bandit modelling is a sequential variation of experimental design, where we also care about our “rewards” as we collect data and perform actions.
Seems Sensible so Far? (I hope)

- Causal inference is not complicated per se, however it does require much attention to detail.

- Crucially, we defined treatment as something “set by design”. What does that mean?

- And isn’t the setting different, you know, when you are actually making decisions later on? How can we generalize?
The Stuff Nightmares are Made Of

The whole complication lies on the definition of “set by design”. We can’t actually formally define it without using causal concepts, and we can’t define causal concepts without the concept of “set by design”.

Introducing: Observational Studies

I used to think correlation implied causation.

Then I took a statistics class. Now I don’t.

Sounds like the class helped. Well, maybe.

Compulsory XKCD strip
Out of Control

- In an observational study, the quantity we deem as the “treatment” is not under any designer’s control.

- Case in point, **smoking** as treatment, **lung cancer** as outcome.

- How would one apply the framework of experimental design to the smoking and lung cancer problem?
Where Do Treatments Come From?

Common causes

Smoking

Lung cancer
Running a Controlled Trial

Randomize

Smoking

Common causes

Lung cancer
Exploiting the Knowledge Learned from a Controlled Trial

Policy

Smoking

Common causes

Lung cancer
Exploiting the Knowledge Learned from a Controlled Trial

- Smoking
- Lung cancer

Common causes
But... We Can’t Randomize

Genetic Profile?

Smoking  Lung cancer
“Adjust”

Genetic Profile

Smoking

? ?

Lung cancer
But... What If?...

Genetic Profile

“Sloppy Lifestyle”? 

Smoking

Lung cancer
And So On

Genetic Profile

"Sloppy Lifestyle"

Aliens?

Smoking

Lung cancer
Observational Studies

- The task of learning causal effects when we do not control the treatment, which instead comes in a “natural regime”, or “observational regime”.

- The aim is to relate use the data in the observational regime to infer effects in the interventional regime.
That Is

We would like to infer \( P(\text{Outcome} \mid \text{Treatment}) \) in a “world” (regime) like this:

All we have is (lousy?) data for \( P(\text{Outcome} \mid \text{Treatment}) \) in a “world” (regime) like this instead:

- Smoking
- Lung cancer
- Common causes
A Historical Example

- Cholera in Soho, 1850s
- Miasma theory: brought by “bad air”
  - No germ theory at the time
- In hindsight: water supply contaminated
- Location was associated with outbreaks
Enter John Snow, “father” of Epidemiology

Here to save the day

Understanding it with Causal Diagrams

• Based on common sense, location was a cause of disease
  • But this didn’t rule out miasma theory

• In one sense, Snow was doing mediation analysis:
  • Location was irrelevant once given the direct cause, water – in particular, one major pump
Understanding it with Causal Diagrams

- Location → Cholera
- Location → Cholera
- Location → Contaminated Water Access → Cholera
- Location → Contaminated Water Access → Cholera
Control, Revisited

• Notice that, in order to maximize a “reward” (minimum expected number of cholera cases), we could have created a policy directly by intervening on Location.

• That is, if you think that “Evacuate Soho for good!” would be a popular policy.

• Mediation matters in practice, and control is more than policy optimization: it is about what can be manipulated in practice or not.
What Now?

- The jump to causal conclusions from observational data requires some “smoothing” assumptions linking different regimes.
A Crude Analogy: Regression, or “Smooth Interpolation”

http://www.gaussianprocess.org/gpml/code/matlab/doc/
What Now?

• To do “smoothing” across regimes, we will rely on some modularity assumptions about the underlying causal processes.

• We just have the perfect tool for the job: Bayesian networks (a.k.a graphical models).
BAYESIAN NETWORKS: A PRIMER
Graphical Models

- Languages for decomposing probabilistic models.

- Because we want sparsity as a means of facilitating estimation and computation.

- But also modularity. We use a graph as a visual representation of a family of factorizations of a probabilistic model.

  - The graph itself is just a drawing: it is the system of constraints encoded by the drawing that is the essence of a graphical model.
  - Vertices are the (random) variables of a probabilistic model.
Bayesian Networks

- A model that follows the structure of a directed acyclic graph (DAG), traditionally for discrete variables.

Task: represent $P(X_1, X_2, X_3, X_4)$
Bayesian Networks

- It is enough to encode the **conditional probability of each vertex given its parents**.

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4) = P(X_1)P(X_2)P(X_3 \mid x_1, x_2)P(X_4 \mid x_2, x_3)
\]
Example: The Alarm Network

Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (□) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume.

Detour: Before Proceeding

- Two simple operations you will need to be familiar with.
- Say you have some $P(X_1, X_2, X_3)$:

  **Marginalization (“sum rule”):**
  
  $$P(X_1 = x_1, X_2 = x_2) = \sum_{x_3} P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

  **Conditioning:**
  
  $$P(X_1 = x_1 \mid X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} \propto P(X_1 = x_1, X_2 = x_2)$$

  **Important**! This is NOT a distribution over $X_2$!
• Factorizations will imply independence constraints. Here, $X_4$ is independent of $X_1$ given $X_2$ and $X_3$

$$P(X_4 = x_4 | X_1 = x_1, X_2 = x_2, X_3 = x_3) \propto P(x_1)P(x_2)P(x_3 | x_1, x_2)P(x_4 | x_2, x_3) \propto P(x_4 | x_2, x_3)$$
Independence Constraints are “Non-Monotonic” in a Bayes Net

\[ P(X_1 = x_1, X_2 = x_2) = \sum_{x_3, x_4} P(x_1)P(x_2)P(x_3 \mid x_1, x_2)P(x_4 \mid x_2, x_3) = P(x_1)P(x_2) \]
Independence Constraints are “Non-Monotonic” in a Bayes Net

\[ P(X_1 = x_1, X_2 = x_2 | X_3 = x_3) \propto P(x_1)P(x_2)P(x_3 | x_1, x_2) \neq g(x_1)h(x_2) \]
Independence Constraints are “Non-Monotonic” in a Bayes Net

\[
P(X_1 = x_1, X_2 = x_2 | X_3 = x_3) \propto P(x_1)P(x_2)P(x_3 | x_1, x_2) \neq g(x_1)h(x_2)
\]

It’s this guy’s fault
Understanding This by “Explaining Away”
The qualitative structure of the system (the graph) allows us to deduce dependencies/independencies which are entailed by it.
Reading Off Independencies

- Conditioning on a “collider” (“v-structure”) activates a path

Conditioning on $X_k$
“activates” $X_i \rightarrow X_k \leftarrow X_j$
Reading Off Independencies

• Conditioning on a “non-collider” **de-activates** (or **blocks**) a path

Conditioning on \(X_k\) “blocks” \(X_i \rightarrow X_k \rightarrow X_p\)
In Our Example

• $X_4$ is independent of $X_1$ given $\{X_2, X_3\}$ because both paths from $X_1$ to $X_4$ are blocked by $\{X_2, X_3\}$
Non-Structural Independencies

- It is possible for some independencies to follow not from the graph, but from particular parameter values.

- This is easier to understand in linear systems.

\[ X_k = b_{i1}X_{p1} + \ldots + b_{ik}X_{pk} + e_k \]
Non-Structural Independencies

• Example

If $b = -ac$, then $X_i$ is independent of $X_{p1}$, even if this is not implied by the graph (and it doesn’t even hold when fixing $X_{p2}$)
What Next

• The decomposition of a system as a graphical model will be the key step to link observational and interventional regimes in the sequel.
FROM GRAPHS TO CAUSAL EFFECTS
Task

• Say you have some **treatment X** and some **outcome Y**.

• Say you have some **background variables Z** you do observe in your data, and which may (or may not) block all paths along common causes of X and Y.

• **Find me a measure of how Y changes when I intervene on X at different levels.**

• But you only have observational data!
Introducing Proper Notation

• For instance, if Y and X are binary, I could be interested in this following **average causal effect**,

\[ P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0) \] under intervention

• **But wait!** This notation can be very confusing. In the observational regime, X is random. In interventional regime, X is fixed by some “magical” agent external to the system.
Introducing Proper Notation

The external agent

Variables ignored by the system
Pearl’s “Do” Notation

- We distinguish random Xs from “fixed” Xs by the notation “do(X)”.

- Average causal effect:

\[ P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0)) \]

- As we say in statistics, this is the estimand. We may derive it from a model, and estimate it with an estimator.

TECHNICAL NOTE: it is still not ideal, as in traditional probability anything to the right of the conditioning bar should be a random variable observed at a particular value. A more kosher notation would be \( P_{do(X = x)}(Y = 1) \) or \( P(Y = 1; do(X = x)) \), but now this has stuck.
PLEASE!

- If you learn one thing from today’s talk, it should be: **do not conflate estimand, with model, with estimator**!

- This is a MAJOR source of confusion, and one of the main reasons why people talk past each other in causal inference.

- Most of my focus will be on clearly defining estimands and models.
The Model

• Now we need a way of deriving this estimand from the observational regime.

• The whole game is to postulate a causal graph, to see how the estimand can be written as a function of it, and to check whether this function can be calculated from the observational regime.
What is a Causal Graph?

- A causal graph is a Bayesian network where the parents of each vertex are its **direct causes**.
What is a Direct Cause?

• The direct causes of $X_i$ are the variables which will change the distribution of $X_i$ as we vary them, as we perfectly intervene in the whole system.

\[
P(X_3 = x_3 \mid \text{do}(X_1 = x_1), \text{do}(X_2 = x_2), \text{do}(X_4 = x_4)) \neq \]
\[
P(X_3 = x_3 \mid \text{do}(X_1 = x_1'), \text{do}(X_2 = x_2), \text{do}(X_4 = x_4))
\]

\[
P(X_3 = x_3 \mid \text{do}(X_1 = x_1), \text{do}(X_2 = x_2), \text{do}(X_4 = x_4)) = \]
\[
P(X_3 = x_3 \mid \text{do}(X_1 = x_1), \text{do}(X_2 = x_2), \text{do}(X_4 = x_4'))
\]
What is a Perfect Intervention?

- A perfect intervention on some $X$ is an independent cause of $X$ that sets it to a particular value, all other things remain equal.
- …
What is a Perfect Intervention?

- We won’t define it. **We will take it as a primitive.**

- “I know it when I see it.”

- Operationally, this just wipes out all edges into X and make it a constant, **all other things remain equal.**

- How is it related to randomization?
Relation to Randomization

- Randomization is NOT a concept used in our definition of causal effect. Nor should it be.
  - Look at the estimand. It is there? No.
- Randomization is a way of sampling data so that we get an estimator that will give a consistent answer.
  - Which is exactly what is missing in an observational study.
  - In practice, if you can do randomization you should.
  - Think of randomization in other contexts, such as estimating public opinion from surveys.
Relation to Other Interventions

• In some cases, we are interested in randomized actions (think of game-theoretical setups, for instance), and/or which might also depend on other variables.

• This just moves the intervention index one level up.

\[
P(Y, X, \text{Common Causes} \mid \text{do}(X' = x'), Z = z)
\]
Another Way of Looking at It

• Graphically, it will be easier to find out what can be learned from observational data if we cast the regime indicator as a single variable, which can be “idle”.

\[
P(Y, X, \text{Common Causes} \mid \text{do}(X = x)) = P(Y, X, \text{Common Causes} \mid F_x = x) \\
P(Y, X, \text{Common Causes} \mid X = x) = P(Y, X, \text{Common Causes} \mid F_x = \text{idle})
\]
Another Way of Looking at It

That is, we will read off independencies that will tell us whether it matters if $F_x$ is “idle” or not.
So, It Boils Down to This (Mostly)

We will try to block pesky hidden common causes to our best.
So, It Boils Down to This (Mostly)

That failing, we will try to exploit some direct causes of the treatment that do not directly affect the outcome.
This is the Bread and Butter of Inferring Causality in Observational Studies
A Starting Example

- Postulated causal graph
A Starting Example

- do(X) regime: module $P(X \mid Z)$ gets replaced by a constant, other modules, $P(Z)$ and $P(Y \mid X, Z)$, remain invariant.

- Can the estimand be derived using observational data only? How?
A Starting Example

- *Ceteris paribus*: we have $P(Y, Z \mid \text{do}(X)) = P(Z)P(Y \mid X, Z)$
- So, straight marginalization gives:

$$P(Y = 1 \mid \text{do}(X = x)) = \sum_z P(Y = 1 \mid X = x, Z = z)P(Z = z)$$
Learning from Data

\[ P(Y = 1 \mid \text{do}(X = x)) = \sum_z P(Y = 1 \mid X = x, Z = z)P(Z = z) \]

- *Now* comes the estimator.
- We can fit a logistic regression to \( P(Y = 1 \mid X = x, Z = z) \) etc. We can fit some kernel density estimator for \( P(Z = z) \) etc. Then plug these estimates in.
• Alternatively, we can fit some \( P(X = x | Z = z) \).
• We can then go through our data points \( \{X^{(i)}, Y^{(i)}, Z^{(i)}\} \) and do the following. Since \( P(Y = 1 | \text{do}(X)) = \mathbb{E}[Y | \text{do}(X)] \),

\[
P(Y = 1 | \text{do}(X = x)) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{1}(X^{(i)} = x)Y^{(i)}}{P(X^{(i)} = x | Z^{(i)} = z^{(i)})}
\]

• This is sometimes called a “model-free” estimator, as it doesn’t fully specify a model.
Learning from Data

- Recall the sum rule

\[
E\left[ \frac{I(X = x)Y}{P(X \mid Z)} \right] = \sum_z P(Y = 1 \mid X = x, Z = z)\frac{P(X = x \mid Z = z)P(Z = z)}{P(X = x \mid Z = z)}
\]

\[
P(Y = 1 \mid \text{do}(X = x)) = \sum_z P(Y = 1 \mid X = x, Z = z)P(Z = z)
\]
Learning from Data

- So it boils down to good models for $P(X \mid Z)$ or $P(Y \mid X, Z)$
- Some methods combine both, so that it allows for some more robust estimation.
Next Example
We will explicitly include the regime indicator $F_x$, such that $P(X = x \mid F_x = \text{idle}, Z) = P(X = x \mid Z = z)$ and $P(X = x \mid F_x = x, Z) = 1$.
Re-arranging It

\[ P(Y | F_x = x) = \sum_z P(Y | F_x = x, Z = z)P(Z = z | F_x = x) \]
Re-arranging It

$$P(Y \mid F_x = x) = \sum_z P(Y \mid F_x = x, Z = z, X = x)P(Z = z)$$

By redundancy

By independence
Re-arranging It

\[
P(Y \mid F_x = x) = \sum_z P(Y \mid Z = z, X = x)P(Z = z)
\]

Identifiable!
Back-door Adjustments

That’s how these types of adjustments are known, and are essentially the backbone of more complex algorithms that can (graphically) answer any possible causal question for a given query.
Next Example
Next Example
Oh, Dear…

We need to condition on H, but we don’t measure it (or aren’t even sure what it is).
Bayes to the Rescue?

Leave this with me and my friends. Gibbs, Metropolis, one of these guys will nail it!
Chances are You Are Going to Screw it Up

More on that later.
Shooting Down a Major Myth

• In practice, researchers try to measure as many possible things that pass as common causes of X and Y as possible, adjust for them, hope for the best.

• Not that I (or anyone) have a universal solution, but this in particular may be a very bad thing to do.
Pearl’s M-bias Example
Shooting Down a Major Myth

• Some researchers in causal inference say this is not very relevant in practice.

• Such comments MIGHT be true-ish for many (which?) practical problems, but they are NOT based in hard evidence or any firm empirical causal knowledge.

• Nobody said causal inference would be easy.
Shooting Down a Major Myth

• Some researchers in causal inference say this is not very relevant in practice.

• Such comments MIGHT be true-ish for many (which?) practical problems, but they are NOT based in hard evidence or any firm empirical causal knowledge.

• Nobody said causal inference would be easy.
A Scary Example

- In linear models with the causal graph below, you are guaranteed to do worse, possibly MUCH worse, by adjusting for Z instead of the empty set.
So, What to Do with this Beast?

- Give up, or
- Try to measure “most” relevant common causes, cross fingers, or
- Look for some external help…
Instrumental Variables

• Say you want to estimate the average causal effect of flu vaccination on health

• Remember: implicit on all examples is the notion your treatments and measurements are well defined.

  • “Vaccination” according to some physical process
  • “Health” as hospitalization in $N$ months from vaccination intake with “flu symptoms”
    • “Flu symptoms” means etc. etc.
In the Wild

• You may have a previous randomized controlled trial (RCT), but the subjects there might differ from the actual population, or the inoculation process changed etc…
An Easier Process to Randomize

• An encouragement design: randomize which physicians receive letters
• Notice the absence of an edge from encouragement to health
Where Does This Take Us to?

• The absence of some edges limits the possible interventional distributions.

• This gives us lower bounds and upper bounds on the causal effect, which may or may not be useful.

• In linear systems it is possible to get the causal effect.
Example: Linear Systems

- With randomized $W$, we assume $W$ and $X$ are correlated.

$$\text{Cov}(W, X) = a \times \text{Var}(W)$$

$$\text{Cov}(W, Y) = a \times b \times \text{Var}(W)$$

So we can get “$b$” out of observational data!

- (Cheeky comment: this is basically “all” of Econometrics)
Non-Linear Systems: Trying to Bayes Your Way Out of It

• Can we get “the” causal effect by latent variable modelling?

• For example, it is not uncommon to conjure latent classes as a way of modelling confounding.

![Diagram](image-url)
Motivation

Bayesian inference is well-defined even in unidentifiable models, so why not?
Do That at Your Own Risk

- Inference is EXTREMELY sensitive to priors.
- Example: binary synthetic data, discrete hidden variable, training data with 1,000,000 points and three different priors.
- Simulation results in next slide.
ACE distribution, mean = -0.29

ACE distribution, mean = -0.05

ACE distribution, mean = -0.07

Silva and Evans (JMLR, to appear)
Alternative Bayesian Inference

OK, alternatively we can define a likelihood function that refers only to observable constraints.
Alternative Bayesian Inference

We can also separate what is identifiable from what is not identifiable for higher transparency.
Example of Analysis: Flu Data

Silva and Evans (NIPS, 2015; JMLR, to appear)
Instrumental Variables and “Broken Experiments”

• Even randomized controlled trials might not be enough.
• Another reason why the machinery of observational studies can be so important.
• Consider the **non-compliance problem** more generally.
Intention-to-Treat and Policy Making

• From the RCT, we can indeed get the intention-to-treat effect.

• From the point of view of policy making, would that be enough?

Diagram:
- Nasty pictures in cigarette packages
- Smoking
- Lung Cancer
- "Risk taking attitude"
A Modern Example

- What is the social influence of an individual or organization?

- It is pointless to define it without causal modelling.
  - Orwellian frame: “If we control the source, we control the followers.”

- Much social influence analysis out there is not necessarily wrong, but it may certainly be naïve.

- Time ordering is very far from enough.
  - Time of measurement is not the same as time of occurrence!
  - What are the common causes?
Broken Experiments of Social Influence

I “like” a particular page

External media exposure

My friend Anna “likes” it a week later
What Facebook-like Companies Would Love to Do

I “like” a particular page

My friend Anna “likes” it a week later

External media exposure
What They Can Actually Do

Expose Ricardo to that Particular Page

I “like” a particular page

External media exposure

My friend Anna “likes” it a week later
Wait, It Gets Worse

- Ricardo’s personality traits
- Anna and I are friends
- Anna’s personality traits
- My friend Anna “likes” it a week later
- External media exposure

I “like” a particular page
Network Data: Possible Solutions

• On top of everything, we need to “de-confound” associations due to the network structure.

• We can of course still try to measure covariates that block back-doors to latent traits.

• Moreover, another compromise is to infer latent variables (stochastic block-models and others), cross fingers, hope for the best.
FROM DATA TO GRAPHS
Adjustments, Causal Systems and Beyond
Those Back-door Adjustments

- Can we get some **proof** or **certificate** we are doing the right thing using data, not only background knowledge?
Structure Learning

- Inferring graphs from testable observations

Graph

\[
X \quad Y
\]

Data

\[
X \perp Y
\]
Structure Learning

- Inferring graphs from testable observations

Data

\[ X \perp\!\!\!\!\!\!\!\!\!\!\perp Y \]

Graph

\[ X \quad Y \]
Structure Learning

- Inferring graphs from testable observations

Data

\[ X \not\equiv Y \]

\[ X \equiv Y \mid Z \]

Graphs (Equivalence class)
Equivalence Class?

- Just life effect identification, graph identification might not be possible. It will depend on which assumptions we are willing to make.

- For instance,
  - Partial ordering
  - Parametric relationships, like linear effects
Main Assumption: **Faithfulness**

- “Non-structural independencies do not happen.”

![Diagram of Truth](image1)

\[
Y \perp Z \\
Y \not\perp Z \mid X
\]

![Diagram of Inference](image2)

\[
Y \perp Z \\
Y \not\perp Z \mid X
\]
• Although in theory “path cancellations” are exceptions, in practice they might be hard to detect.

• However, faithfulness can be a very useful tool for generating models compatible with the data that you actually have. Taking other people’s theoretical graphs at face value is unnecessary.

• Other default alternatives, like “adjust for everything”, are not really justifiable. You should really try a whole set of different tools.
Example

• W not caused by Y nor Y, assume ordering X → Y
• W ⊥⊥ X, W ⊥⊥ Y | X + Faithfulness. Conclusion?

Naïve estimation works:
Causal effect = P(Y = 1 | X = 1) − P(Y = 1 | X = 0)

• This super-simple nugget of causal information has found some practical uses on large-scale problems.
Application

- Consider “the genotype at a fixed locus $L$ is a random variable, whose random outcome occurs before and independently from the subsequently measured expression values”

- Find genes $T_i, T_j$ such that $L \rightarrow T_i \rightarrow T_j$

Chen, Emmert-Streib and Storey (2007) Genome Biology, 8:R219
Validating or Discovering Back-door Adjustments

  - Generalizes “chain models” $W \rightarrow X \rightarrow Y$

R1: If there exists a variable $w \in W$ and a set $Z \subseteq W \setminus \{w\}$ such that

(i) $w \perp y \mid Z$, and
(ii) $w \perp y \mid Z \cup \{x\}$

then infer ‘±’ and give $Z$ as an admissible set.
Illustration

R1: If there exists a variable \( w \in \mathcal{W} \) and a set \( \mathcal{Z} \subseteq \mathcal{W} \setminus \{w\} \) such that

(i) \( w \not\perp y \mid \mathcal{Z} \), and

(ii) \( w \perp y \mid \mathcal{Z} \cup \{x\} \)

then infer ‘\( \pm \)’ and give \( \mathcal{Z} \) as an admissible set.

- Notice the link to instrumental variables.
System-Wide Causal Discovery

- Finding the graph for a whole system of variables
System-Wide Causal Discovery

- Equivalence class: one edge fully unveiled.

Combining Experimental and Observational Data

AVOIDING MINE TRAPS

Think through your problem, don’t just big data a solution out of it.
Don’t Take Your Measurements and Interventions for Granted
What Does That Mean?

Common Causes

Gender

Hiring
What About This?

I’d Settle on This

Gender

Common Causes

Hiring

Blind auditions

Gender Perception
More Controversially, What about Innate Effects in the Example?

- I’d appeal to Faithfulness and see how Gender and Hiring can be made independent by Gender Perception and other covariates.
But What Does That Mean???

- Gender
- Perceptions
- Hiring

Blind auditions

Gender

Common Causes

Gender Perception

Hiring
Ideal Interventions, Again

• Some researchers believe that if there is no physically well-defined mechanism of intervention, then the causal question should not be asked.

• I believe the above is non-sense.
  
  • Do genders have different effects on particular diseases?
  
  • What about disentangling whether being male leads to higher rates of heart attacks, or whether this is just confounded by behavioural effects or other genes. Why wouldn’t we want to ask these questions?

• See Pearl (2009) for more on that, which is a primary defence of ideal interventions. But this is NOT a license for not paying attention to what your variables mean.
Regression and Causation

- It has been trendy for a while to fit big regression models and try to say something about “variable importance”.
- Again, what does that mean?
- If you want to make causal claims, **say it**, don’t pretend this is not your goal.

Fantasy

```
X_1 -----> Y
     |      |
     V      V
     X_2
```

Reality?

```
H_1
    /  \
   /    \
X_1   X_2
      /  \
     /    \
    H_2
```
Conditioning and/or Intervening: What is that that You Want?

The “paradox”:

\[ P(E | F, C) < P(E | F, \neg C) \]
\[ P(E | \neg F, C) < P(E | \neg F, \neg C) \]
\[ P(E | C) > P(E | \neg C) \]

Which table to use? (i.e., condition on gender or not?)

(Pearl, 2000)
Some Possible Causal Graphs

(a) Treatment $C$ → Gender $F$ → Recovery $E$

(b) Treatment $C$ → Blood pressure $F$ → Recovery $E$

(c) Treatment $C$ → Recovery $E$
Dissolving a Paradox Using Explicit Causal Modelling

• Let our population have some subpopulations
  • Say, F and \( \sim F \)

• Let our treatment C not cause changes in the distribution of the subpopulations
  • \( P(F \mid \text{do}(C)) = P(F \mid \text{do}(\sim C)) = P(F) \)

• Then for outcome E it is impossible that we have, simultaneously,
  • \( P(E \mid \text{do}(C), F) < P(E \mid \text{do}(\sim C), F) \)
  • \( P(E \mid \text{do}(C), \sim F) < P(E \mid \text{do}(\sim C), \sim F) \)
  • \( P(E \mid \text{do}(C)) > P(E \mid \text{do}(\sim C)) \)
Proof

\[
P(E|do(C)) = P(E|do(C), F)P(F|do(C)) + P(E|do(C), \neg F)P(\neg F|do(C))
\]
\[= P(E|do(C), F)P(F) + P(E|do(C), \neg F)P(\neg F).
\]

\[
P(E|do(\neg C)) = P(E|do(\neg C), F)P(F) + P(E|do(\neg C))P(\neg F)
\]

\[
P(E|do(C)) < P(E|do(\neg C)),
\]
CONCLUSIONS
Yes, It is Hard, But:

- Pretending the problems don’t exist won’t make them go away.

- There is a world out there to better explored by combining experimental and observational data.

- In particular, how to “design experimental design”.

- The upside of many causal inference problems is that getting lower bounds and relative effects instead of absolute effects might be good enough.
Main Advice

Don’t rely on a single tool. If you can derive similar causal effects from different sets of assumptions, great. If they contradict each other, this is useful to know too. Make use of your background knowledge to disentangle the mess.
Textbooks


http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/

Excellent, but be warned: verbose
Classics For Researchers


Let us Let Fisher Have the Last Word

5. The Step from Association to Causation

This issue is naturally of great concern to workers in observational research and has received much discussion in individual subject-matter fields. I shall confine myself to a few comments on statistical aspects of the problem.

First, as regards planning. About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: “Make your theories elaborate”. The reply puzzled me at first, since by Occam’s razor the advice usually given is to make theories as simple as is consistent with the known data. What Sir Ronald meant, as the subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold. If a

The Planning of Observational Studies of Human Populations
W. G. Cochran and S. Paul Chambers
Journal of the Royal Statistical Society. Series A (General)
Vol. 128, No. 2 (1965), pp. 234-266
Or Maybe Not. Thank You

“I’d rather have another beer now than be Fisher.”